Appendix 1

Validity and Reliability of PATB

Studies of Validity and Reliability of PATB

To judge the adequacy of the evidence relating to the validity and reliability of PATB I and II, we reviewed 23 studies that had been done either by or under the auspices of the staff of PSS and the Test Data Book No. 15, published 1 July 1958 and had a one day conference with the Chief and the staff of PSS.

STAT/INTL

In our search for evidence on the validity and reliability of PATB, we discovered that not all applicants for professional positions in the Agency were required to take the test battery. We could find no written policies as to who was or was not required to take the tests. Some of the exceptions, such as applicants for positions that require highly specialized knowledge or competencies that are not tested by PATB or applicants who were specifically recruited because they were known to have expert knowledge in an area of high priority to the Agency, seemed reasonable. However, it did not seem reasonable that among candidates with similar educational and experience backgrounds applying for the same job in the same unit, some were required to take PATB and some were not. We were not able to determine precisely how many applicants for professional positions in

the different units of the Agency were required to take PATB before the decision to employ them was made. Our best estimate of the number was obtained from the OIG survey of entrants on duty from 1 October 1977 to 31 August 1979. This estimate is based on self-reports of the EOD's and is subject to the inaccuracies of such reports. According to the survey data, there were 218 EOD's of professional status; of these, 137 (63%) reported that they had taken PATB and 81 (37%) reported that they had not taken it. We were not able to determine whether female and minority applicants for professional positions were required to take PATB more frequently than white males to be considered for employment. This should be investigated because if it proves to be true, then the policy violates the Uniform Guidelines on Employee Selection Procedures[1/] (1978). Whether it is true or not, we recommend that the Agency develop a systematic policy on personnel selection policies and the role of PATB in personnel selection.

1/ Equal Employment Opportunity Commission. Uniform guidelines on employment selection procedures. Federal Register, August 25, 1978, 43 (166) p. 38300, Sec. 11

During our investigations, we also discovered that some units in the Agency were using tests either devised in the unit or taken from other sources for selection of personnel. We were unable to discover the full extent of the practice nor did we find any research on the validity or reliability of these tests. Although our assignment was to examine the validity and reliability of PATB, we think we should bring to the attention of the Agency the need to validate all tests and procedures used for the selection of personnel and also the need to control the use of unvalidated personnel selection procedures.

Before we present our evaluation of the evidence for validity and reliability of PATB, one additional point needs to be made. All of the materials that were made available to us focused on the scores on the separate tests and their relationships to criterion ratings of job performance. However, the people in the units of the Agency who are responsible for making the decisions about employment never see specific scores on the tests. Instead, they see a narrative report written by a staff member in PSS in which an individual's performance on the cognitive tests is reported in adjectival categories ranging from very poor

to very superior. In addition, the narrative report contains a "personality description" based on the work attitude scales and the temperament scales, a description of vocational interests, an adjectival rating of foreign language proficiency, a somewhat general description of writing skills, and usually a recommendation as to the desirability of employing the candidate or for employment in a particular unit or job category.

The narrative report is partly description, partly prediction, and partly fantasy. If performance on the PATB influences the decision to hire or not to hire an applicant, it can do so only through the narrative report. Under these conditions, it is the validity of this narrative report that is central, and its validity basically depends upon the adequacy of the evidence for the content, construct and criterion-related validity of the individual tests that are included in PATB I and II.

How is this supported?

Content and Construct Validity of PATB

PATB was constructed and implemented in the 1950's. In the initial construction of a test battery to be used for selection of personnel, the constructor of the battery

should first attempt to establish the content and construct validity of each test in the battery. A test is considered to have content validity when the tasks called for by the test closely match tasks done on the job. A good illustration would be a typing test in which the type of copy corresponds to that which the typist would be called upon to work upon after employment. A test is considered to have construct validity when the attributes called for by the test are similar to those called for by essential job tasks. By way of illustration, if a job required the incumbent to draw appropriate conclusions from complex sets of data, then test exercises that presented data of various types and called upon the examinee to determine which of a series of conclusions followed from the data would appear to have construct validity for that job.

Determination of content and construct validity is primarily a rational process. Its foundation is a detailed, analytic study of the job to determine what the incumbent is in fact called upon to do. When the tasks that occur frequently or are of critical importance have been determined, the job analyst must try to judge what competencies are required if the incumbent is to perform these jobs

effectively. On the other side, the job analyst must examine existing or proposed tests and try to judge what competencies are required in order to get good scores on these tests. Certain types of statistical analysis of test results are helpful in gaining an understanding of tests. These tend to be studies of what other tests the test in question does or does not correlate with. For example, a test of mechanical ability would be suspect if it showed too high a correlation with a vocabulary test. It might then be thought to be too highly contaminated with verbal ability. One formal approach to the analysis of test relationships that is often instructive is that of factor analysis, in which one tries to identify and define the factors entering into each test in a battery of tests. The C/PSS reported that a factor analysis had been done of the battery but, for some unexplained reason, it had not been kept.

In our review of the PATB and its use, we inquired to determine what information existed on the tasks done in and competencies required for Agency jobs. We found very little of a systematic nature. Brief and rather general statements have been prepared for distribution to recruiters. We are

not able to judge how thorough an analysis lies back of these descriptions, but we found no evidence that they were based on careful and detailed study. Members of the Psychological Services Staff report that they do discuss the job to be filled with persons in the different branches of the Agency, and do acquire a personal and internal impression of what the job is like and what competencies are called for. However, these impressions are not formalized in any record or systematically reviewed for accuracy.

Though anyone who looks at the recruiter guides and talks with Agency personnel gets some impression of what demands the different jobs in the Agency make, these impressions tend to be quite subjective and superficial. They are unsystematic, undocumented and unrecorded. There is nothing that would stand up to critical scrutiny, or that would meet the standards of the APA[1/] or of the EEOC[2/]. The available information provides, at best, weak support for the validity of the present battery, and little basis for improving it.

---

1. Standards for Educational and Psychological Tests. Washington, D.C. American Psychological Association, 1974.

2. Equal Employment Opportunity Commission (EEOC) Uniform guidelines on employee selection procedures. Federal Register, August 25, 1978, 43 (166), 38290-38315.

We found no statement of the rationale for the structure of the current battery. If one ever existed, it has apparently disappeared in the mists of antiquity. It would appear, even to the casual observer of the work of the Agency, that many of the jobs do make demands for a high level of ability on cognitive functions of obtaining, synthesizing, and processing information. Since tests of reading comprehension, arithmetical reasoning, abstract reasoning, writing and data interpretation do appear to call for some of these cognitive abilities, there seems to be a fair congruence between the functions measured by these tests and the demands made by a number of the Agency jobs. However, this is a superficial reaction. Whether the fit is as close as it could be, whether all the tests do in fact relate to job demands, and whether the range of job demands is covered as completely as possible by the present battery is something that we think cannot be answered without a substantially more detailed and penetrating analysis of job requirements. We believe that evaluation of the present battery and steps toward its improvement must rest on such analyses.

Criterion-Related Validity

Since there is no adequate evidence of the content or construct validity of PATB, continued use of the battery

depends entirely upon the quality of evidence for criterion-related validity. As contrasted with content or construct validity which depend primarily upon logical analysis, criterion-related validity is always statistical and depends upon determining whether performance on the tests in PATB is significantly related to measures of job performance.

Problems in Determining Criterion-Related Validity

For criterion-related validation studies of selection tests to be fully satisfactory, it is necessary that

(1) test data be available for a sample of substantial size working in the same or closely similar jobs,

(2) the sample working in the job be representative of the pool of applicants for the job, or if a select group has been admitted, the exact basis and extent of that selectivity be known, and

(3) relevant, reliable and unbiased indicators of job success be available for each person working in the job.

We tried to find out how well these requirements are approximated in the Agency. One very useful source of information was an analysis carried out for us by the Psychological Services Staff of the 2359 persons who had been tested with PATB, Part I, during 1978. Of these, only 222 (9.4%) were identified as having been actually employed

by the Agency. These persons were spread over some 30 different job categories, only 3 of which contained as many as 20 persons. The categories containing as many as 5 cases were the following:

Job Category	No. of Cases
Career Trainee (CMS)	55
Intelligence Officer	40
NPIC Specialist	24
STATSPEC [redacted]	15
Office of Security Investigator	14
Career Trainee (not CMS)	9
Programmer	7
Operations Officer	6

These figures suggest that validation studies could be meaningfully carried out for at most 2 or 3 specialties, and for these only if the results from two or more years of testing were combined.

The need for large samples is much enhanced when a number of different scores are all validated at the same time in a shot-gun approach. This has been typical of validation studies in the Agency. The PATB yields some 30 separate scores when one counts all the separate scores from the temperament survey and the work attitudes questionnaire.

Just as one will just by chance occasionally hit a run of 10 reds in a row on the roulette wheel, so one will by chance occasionally get the appearance of validity for a test in a sample of cases. The chances are greatly increased when a whole collection of different scores is studied. Under these circumstances, it is crucial that one cross-validate any apparently valid test, i.e., that one verify the validity in a new independent sample. This requires that a new group of employees be available to study in the job, and in many less common jobs the new sample may simply not exist. It is worth noting that we found almost no instances of cross-validation in the studies carried out in the Agency.

because Ni are so limited

For effective criterion-related validation studies, it is important that those employed on the job be representative of the population of job applicants, or that one knows on what basis any selection has taken place. If those likely to fail or be inferior on the job have been screened out by effective selection procedures, it will be impossible to get a complete picture of the relationship of predictor measures to job success. Those that the predictor would have identified as potential failures may never have been employed.

The study of 1978 examinees was carried out specifically to throw light on this question. To what extent are those

scoring low on the tests screened out so that they never have a chance to appear in later validation studies? We will illustrate the situation here with the Intellectual Composite (sum of standard scores on four cognitive tests), but the effect appears in varying degrees in most of the cognitive tests taken singly. The range of scores obtained on the composite was divided into five equal segments: very low (0-7), below average (8-14), average (15-21), above average (22-28) and high (29-35). Figure 1 shows what happened to those in each of the five groups. One can see that far too few of the below average and low groups were hired to permit any meaningful study of success on the job in these groups. The selectivity may well have worked to the advantage of the Agency, but it tends to be disastrous for validation research. Results for specific measures are summarized in Table 1, for consideration by anyone who is interested in further study. Clearly, effects of the selection procedures have been most marked in the case of the tests of cognitive abilities which probably accounts, at least partially, for the failure to find systematic significant relationships among scores on the cognitive tests and ratings of job performance.

Figure 1

Disposition of Applicants Scoring at Different Levels on the Intellectual Composite from PATB I

Score on Intellectual Composite	Disposition
High (29-35) (Total N = 120)	23 took PATB II and were hired.(19.2%) 56 took PATB II but were not hired 41 took only PATB I, not hired
Above Average (22-28) (Total N = 508)	77 took PATB II and were hired (15.2%) 241 took PATB II but were not hired 190 took only PATB I, not hired
Average (15-21) (Total N = 893)	79 took PATB II and were hired (8.8%) 434 took PATB II but were not hired 380 took only PATB I, not hired
Below Average (8-14) (Total N = 573)	26 took PATB II and were hired (4.5%) 295 took PATB II but were not hired 252 took only PATB I, not hired
Low (0-7) (Total N = 202)	3 took PATB II and were hired (1.5%) 87 took PATB II but were not hired 112 took only PATB I, not hired

Table 1

Comparison of Applicants with Hired Group

	Mean Score		Increase in Employed (in hundredths of S.D.)	Standard Deviation		Percent Reduction in Variance
	Appl.	Emp.		Appl.	Emp	
Intellectual Composite	17.30	21.00	53	6.89	6.18	21
Figure Matrices	4.69	5.52	38	2.20	2.00	17
Reading Vocabulary	3.70	4.77	50	2.15	2.02	11
Reading Comprehension	4.16	5.26	49	2.24	2.02	19
Arithmetic Problems	4.80	5.43	29	2.14	2.04	8
Contemporary Affairs	3.93	4.69	35	2.18	2.00	16
Interpretation of Data	4.08	4.70	30	2.04	1.96	7
Considerations	4.54	4.82	13	2.14	2.17	*
Numerical Operations	3.89	4.19	14	2.10	2.01	8

Table 1 Continued

	Mean Score Appl.	Mean Score Emp.	Increase in Employed (in hundredths of S.D.)	Standard Deviation Appl.	Standard Deviation Emp.	Percent Reduction in Variance
Temperament						
Quick	4.41	4.48	4	1.79	1.91	*
Physical	5.16	5.24	4	2.00	1.94	6
Outgoing	4.54	4.71	9	1.97	1.88	9
Predominant	5.35	5.65	14	2.08	1.99	9
Confident	4.71	4.92	12	1.73	1.71	2
Solitary	4.83	4.71	-7	1.80	1.71	10
Question	5.09	4.71	-20	1.80	1.76	13
Work Attitudes (low score is favorable)						
Training	3.92	3.72	11	1.79	1.50	30
Hazards	3.58	3.60	-1	2.01	1.92	9

Table 1 Continued

	Mean Score Appl.	Mean Score Emp.	Increase in Employed (in hundredths of S.D.)	Standard Deviation Appl.	Standard Deviation Emp.	Percent Reduction in Variance
Analyze	4.05	3.73	15	2.16	2.00	14
Annoyances	4.12	3.98	6	2.23	2.16	5
Reward	3.99	3.93	3	2.19	2.12	6
Soc. Resp.	3.92	3.85	3	2.08	2.12	*
Mechanical	3.46	3.36	6	1.82	1.74	9
Supervisor	4.18	3.90	14	1.95	1.81	14
Physical	4.50	4.11	14	2.83	2.95	*
Supervisee	3.61	3.69	-4	1.92	2.00	*
Soc. Deprivation	3.96	4.12	-8	1.94	1.84	10
Undesirables	4.02	4.07	-2	2.00	2.17	*
Resourcefulness	4.19	3.75	23	1.89	1.82	7
Security/ Unconvent.	5.33	5.32	0	2.09	1.84	22
Temp.	4.69	4.53	8	1.94	1.93	2

It is not clear how the selection evident in Figure 1 and Table 1 came about. The test scores are reflected, though not uniformly, in the narrative reports prepared by the Psychological Services Staff, and in some cases these reports will have influenced the hiring decision of the person in the Agency called upon to make that decision. But cognitive test performance is also related to prior academic record and to educational institution attended. Selectivity based on these factors would have had an indirect effect upon the range of test scores among employees. We simply do not know through what channels test scores were related to employment decisions, and hence we can make no sound adjustment for the selectivity that has occurred.

In general, we must conclude that criterion-related validation studies will be seriously hampered by the very real, but largely unanalyzable, selectivity that has intervened between being an applicant and becoming an employee.

The third major problem area in criterion-related validation studies lies in the short-comings of available criterion indicators of job performance. In most of the Agency jobs there can, by the nature of the job, be no objective record of job performance. One is of necessity

thrown back upon some type of rating by a supervisor. Personnel psychologists have struggled to overcome the subjectivity, rater idiosyncracy and low reliability of supervisory ratings for 60 years but with limited success. There is no reason to believe that conditions in the Agency are any more favorable than elsewhere for obtaining reliable and unbiased judgments from supervisory personnel. So even if larger and fully representative samples of applicants were available in each of the job categories of interest, criterion-related validation studies would still be severely limited.

Critique of Criterion-Related Studies Done in Agency

There were 23 studies made available to us that presumably dealt with the criterion-related validity of PATB. Of these, 4 were studies of success in foreign language training courses and will be discussed in a separate section. Two of

STATINTL

[redacted]

not report any data on PATB. Six of the studies had to be discarded because of inadequate reporting of data. These 6 studies reported no means, no standard deviations, and no correlations with criterion measures

STATINTL

[redacted]

compared a sample of black professional employees and a matched sample of white professional employees to determine whether PATB had differential validity for minority persons. This study had to be discarded because of inappropriate statistical analyses and inadequate reporting of the data.

The Test Data Book No. 15 (1 July 1958) reported criterion-related validity data for each test in the battery

but almost all of the data were against training criteria, not job performance criteria. Validity data for training cannot be used to demonstrate job related validity; therefore these data, too, had to be discarded.

Elimination of these sources left only 10 studies that were judged to meet at least minimal standards for a technical report. In each of these 10 studies, the investigators invested considerable time and effort in obtaining reliable criterion ratings of job performance from supervisors. Most of the studies (7 out of 10) used multiple criterion measures. Despite the care taken to obtain good criterion measures, there was evidence in most of the studies indicating that the ratings given by supervisors were greatly influenced by length of employment and GS level which were irrelevant to the criterion measures. In the studies that used multiple criterion measures, the different measures tended to be highly correlated indicating that there was a general halo effect operating.

All of the 10 studies suffered from the problems that were pointed out in the previous section; namely, small sample size and probably some restriction in range of scores on the tests. The number of subjects used in each of the studies was extremely small in relation to the number of variables used. The number of subjects in the studies

ranged from 10 to 138 whereas the number of variables used ranged from 31 to 510. A rule of thumb for determining the number of subjects needed for multiple regression or discriminant analysis is 10 subjects for each variable. Using this rule of thumb, it is quite clear that none of the studies had an adequate number of subjects for the analyses that were done. With the limited number of subjects available for each study, it is not surprising to find that in none of the studies did the number of significant correlations exceed a chance level.

Only 4 of the 10 studies reported both means and standard deviations for each of the tests for the subjects used. In these studies the standard deviations of the test scores did not differ markedly or in any systematic way from those for the group on which the tests were originally normed. However, the original normative group was made up of persons already employed by the Agency and were probably a more select group than the applicant group. Because of the inadequacies of reporting of data in these studies, it is impossible to determine with any degree of precision how much restriction in range of scores there really was in the employed groups. There is no doubt that there was some and

that the size of the correlations between the test scores and job performance criteria was somewhat limited by the restriction in range.

We have summarized the results reported in the 10 studies for the cognitive tests, work attitude, and temperament scales, in Table 2 and Table 3. Table 2 shows the number of significant correlations found for 5 analyst type jobs in the Agency and Table 3 shows the same information for 6 other job categories in the Agency. Both Table 2 and Table 3 clearly show that the number of significant correlations between scores on the separate tests of PATB and criterion measures of job performance does not exceed a chance level. The lack of consistency of correlations of individual test scores with job performance measures across similar jobs is also an indication that one is dealing with random, rather than true, relationships.

Inspection of Table 2 which summarizes the studies done on analyst type jobs can be used to illustrate the lack of any consistency in the pattern of correlations. The 1967 study of [redacted] was designed to be a cross-validation of the 1958 study of [redacted] The 1967 study found no significant correlations between the scores on the individual tests and any rating of job performance. The findings for

STATINTL
STATINTL

Table 2

Number of Correlations
Significant At .05 Level
Or Better For Analyst Type
Of Jobs In The Agency

	Date	1974	1967	1965	1958	1973
STATSPEC	Group Studied	[redacted]	MRA	MEB	ERA 1/	OSR 2/
	N	N=10	N=35	N=11 to 18	N=40	N=138
	Tests					
	F.M.	0/10	0/1	0/1	4/5	1/5
	R.V.	3/10	0/1	0/1	0/5	-1/5
	R.C.	0/10	0/1	0/1	1/5	0/5
	CAT	2/10	0/1	0/1	0/5	0/5
	AP	0/10	0/1	1/1	0/5	1/5
	IDY	0/10	0/1	0/1	1/5	1/5
	Con	0/10	0/1	-1/1	0/5	-1/5
	N.O.	0/10	0/1	0/1	0/5	0/5
	A.L.	1/10 -1/10	-	-	-	-
	WA Training	0/10	0/1	0/1	0/5	0/5
	WA Hazards	0/10	0/1	-1/1	0/5	2/5
	WA Analyze	0/10	0/1	0/1	1/5	1/5
	WA Annoyances	0/10	0/1	0/1	0/5	2/5
	WA Rewards	0/10	0/1	0/1	2/5	0/5
	WA Soc. Resp.	-2/10	0/1	0/1	1/5	0/5
	WA Mech.	0/10	0/1	0/1	0/5	0/5
	WA Supervisor	0/10	0/1	0/1	0/5	2/5
	WA Phys. Dem.	0/10	0/1	-1/1	0/5	0/5
	WA Supervisee	0/10	0/1	0/1	1/5	0/5
	WA Soc. Depr.	-1/10	0/1	0/1	0/5	0/5
	WA Undesirables	0/10	0/1	0/1	1/5	-1/5
	WA Resourceful	-1/10	0/1	-1/1	0/5	2/5
	WA Tempo	0/10	0/1	0/1	0/5	0/5

Table 2 Continued

Date	1974	1967	1965	1958	1973
STATSPEC Group Studied	[redacted]	MRA	MEB	ERA 1/	OSR 2/
N	N=10	N=35	N=11 to 18	N=40	N=138
Tests					
TTS Quick	0/10	0/1	None Reported	0/5	0/5
TTS Physical	0/10	0/1		0/5	0/5
TTS Outgoing	0/10	0/1		0/5	0/5
TTS Predominant	4/10	0/1		0/5	0/5
TTS Self-Conf.	0/10	0/1		0/5	1/5
TTS Solitary	-1/10	0/1		0/5	0/5
TTS Question	0/10	0/1		0/5	0/5

1/ Used 14 criterion measures but reported on only 5.

2/ Used 8 criterion measures but reported on only 5.

Note: Figures in body of table show number of significant correlation over number of criterion measures used. For example, for column 1, FM, 0/10 means no significant correlations between figure matrices test and any of the 10 criterion measures used in study.

A - sign before a number indicates a significant negative correlation. No sign indicates a positive correlation.

the 1974 study of [redacted] should be disregarded because of the extremely small sample size of 10. The inconsistency of the correlational patterns is also indicated by the fact that scores on the same test or scale correlate positively in one small sample and negatively in another small sample. This is especially true for scores on the work attitude scales. For example, Work Attitudes Hazards and Work Attitudes Resourcefulness correlate negatively; i.e., low scores on these scales are related to high ratings of job performance for MEB analysts but correlate positively with job performance ratings of OSR analysts; i.e. high scores on these scales are related to high ratings of job performance. The inconsistent direction of relationships makes no logical or psychological sense.

STATSPEC

In Table 3, there are two studies reported for data processors, one done in 1969 and one in 1978 (see columns 7 and 8). The 1978 study could be considered a replication of the 1969 study and comparison of the two shows some consistency of relationships between the scores on some of the individual tests and some of the ratings of job performance. The six tests that show some consistency of relationship are Figure Matrices, Reading Vocabulary, Reading Comprehension, Arithmetic Problems, Interpretation of Data and Work Attitudes Training. None of the other cognitive tests, work

Table 3
Number of Correlations
Significant at .05 Level Or
Better For Various Job Categories

Date	1 1973	2 1974	3 1956	4 1956	5 1974	6 1974	7 1969	8 1978
Group Studied	DCD IO	CT's	Cable Analyst	Cable Analyst Trainee	[redacted]		Data Processors	Data Processors
N	N=25	N=70	N=18	N=16	N=51	N=42	N=18 to 46	N=55 to 77
Tests								
F.M.	0/17	0/4	0/1	0/1	0/10	0/10	1/5	5/5
R.V.	-1/17	0/4	0/1	0/1	1/10	0/10	2/5	3/5
R.C.	-7/17	-1/4	0/1	1/1	0/10	-3/10	4/5	3/5
CAT	1/17	1/4	0/1	0/1	0/10	5/10	0/5	0/5
AP	3/17	1/4	0/1	0/1	0/10	0/10	1/5	5/5
IDY	2/17	2/4	0/1	0/1	0/10	0/10	2/5	3/5
Con	0/17	0/4	0/1	1/1	1/10	-2/10	-1/5	1/5
N.O.	0/17	0/4	1/1	0/1	0/10	0/10	0/5	3/5
A.L.	-2/17	0/4	0/1	0/1	0/10	0/10	-	3/5
WA Training	1/17	0/4	0/1	-1/1	0/10	0/10	-1/5	-1/5
WA Hazards	0/17	3/4	0/1	0/1	0/10	-1/10	0/5	0/5
WA Analyze	1/17	0/4	0/1	-1/1	0/10	-1/10	0/5	0/5
WA Annoyances	0/17	2/4	0/1	0/1	0/10	0/10	0/5	0/5
WA Rewards	0/17	0/4	0/1	0/1	0/10	0/10	0/5	0/5
WA Soc. Resp.	0/17	0/4	0/1	0/1	0/10	0/10	-1/5	0/5
WA Mech.	6/17	0/4	0/1	0/1	1/10	2/10	0/5	-1/5
WA Supervisor	0/17	0/4	0/1	0/1	0/10	0/10	-1/5	0/5
WA Phys. Dem.	2/17	0/4	-1/1	0/1	0/10	0/10	0/5	-1/5
WA Supervisee	0/17	0/4	0/1	0/1	0/10	0/10	0/5	0/5
WA Soc. Depr.	0/17	1/4	0/1	0/1	0/10	0/10	-2/5	0/5
WA Undesirables	6/17	0/4	0/1	0/1	0/10	0/10	0/5	0/5
WA Resourceful	0/17	0/4	0/1	0/1	0/10	0/10	0/5	-1/5
WA Tempo	5/17	0/4	-1/1	0/1	0/10	0/10	0/5	0/5

Table 3 Continued

	1	2	3	4	5	6	7	8
Date	1973	1974	1956	1956	1974	1974	1969	1978
Group Studied	DCD IO	CT's	Cable Analyst	Cable Analyst Trainee	[redacted]		Data Processing	STATSPEC Data Processing
N	N=25	N=70	N=18	N=16	N=51	N=42	N=18 to 46	N=55 to 77
Tests								
TTS Quick	1/17	-1/4	0/1	0/1	-1/10	0/10	0/5	0/5
TTS Physical	10/17	0/4	0/1	0/1	0/10	-2/10	0/5	0/5
TTS Outgoing	0/17	0/4	0/1	0/1	0/10	0/10	1/5	0/5
TTS Predominant	0/17	0/4	0/1	0/1	0/10	0/10	0/5	1/5
TTS Self-Conf.	3/17	-2/4	0/1	0/1	0/10	0/10	0/5	0/5
TTS Solitary	-1/17	0/4	0/1	0/1	0/10	0/10	0/5	0/5
TTS Question	0/17	0/4	0/1	0/1	0/10	0/10	0/5	0/5

Note: Figures in body of table show number of significant correlations over the number of criterion measures used. For example, in column 1 for FM, 0/17 means no significant correlations between Figure Matrices test and any of the 17 criterion measures used in the study. A - sign before a number indicates a significant negative correlation. No sign indicates a positive correlation.

attitude or temperament scales shows any consistent correlations with rated job performance for data processors. Among the other five job categories shown in Table 3, the relationship among scores from PATB and the criterion measures of job performance are less consistent. What little consistency there was appeared in the cognitive tests. There was no consistency of relationship between scores on the work attitude and temperament scales and ratings of job performance.

As we were examining the 10 studies we noted that several investigators had used identical or highly similar criterion measures of job performance. Since our analysis of the overall results of the validity studies shown in Tables 2 and 3 revealed extremely weak and unconvincing evidence for the criterion-related validity of PATB, we decided to look at these common criteria across job categories to determine whether scores on PATB would show consistent significant correlation with specific criteria of job performance. The results of the analyses are presented in Table 4. The first four criterion measures--writing ability, oral communication, substantive knowledge and ability to conduct analysis research-- were used in 4 or 5 studies and would be expected to involve largely cognitive abilities.

Table 4

Number of Significant Correlations Among Scores on PATB and Specific Criterion Measures That Were Common to Four or More Studies

Criterion Measure	Writing Ability	Oral Comm.	Substantive Knowledge	Ability to Conduct Analysis Res.	Independence Initiative
Number of Studies	4	5	4	5	4
Tests.					
F.M.	1/4	0/5	0/4	1/5	0/4
R.V.	1/4	2/5	0/4	0/5	0/4
R.C.	-1/4	-1/5	0/4	0/5	0/4
CAT	0/4	0/5	3/4	1/5	0/4
AP	0/4	0/5	0/4	0/5	0/4
IDY	0/4	0/5	1/4	0/5	0/4
Con	0/4	0/5	0/4	0/5	0/4
N.O.	0/4	0/5	1/4	0/5	0/4
A.L.	-				
WA Training	0/4	0/5	0/4	0/5	0/4
WA Hazards	0/4	0/5	0/4	0/5	1/4
WA Analyze	2/4	0/5	0/4	0/5	1/4
WA Annoyances	0/4	0/5	0/4	0/5	1/4
WA Rewards	1/4	0/5	0/4	1/5	0/4
WA Soc. Resp.	0/4	-1/5	0/4	0/5	-1/4
WA Mech.	1/4	0/5	1/4	0/5	1/4
WA Supervisor	0/4	0/5	0/4	1/5	1/4
WA Phys. Dem.	1/4	0/5	0/4	0/5	0/4
WA Supervisee	0/4	0/5	0/4	0/5	0/4
WA Soc. Depr.	0/4	0/5	0/4	0/5	0/4
WA Undesirables	1/4	0/5	0/4	1/5	1/4
WA Resourceful	0/4	0/5	0/4	0/5	1/4
WA Tempo	0/4	0/5	0/4	0/5	1/4
TTS Quick	0/4	0/5	-1/4	0/5	0/4
TTS Physical	1/4	-1/5	-1/4	0/5	-1/4
TTS Outgoing	0/4	0/5	0/4	0/5	0/4
TTS Predominant	0/4	1/5	0/4	0/5	2/4
TTS Self-Conf.	0/4	0/5	1/4	0/5	0/4
TTS Solitary	0/4	0/5	0/4	0/5	0/4
TTS Question	0/4	0/5	0/4	0/5	0/4

Table 4 Continued

Criterion Measure	Adaptability Flexibility	Supervisory Skill Managerial Potential	Organizational Ability	Overall Ratings of Job Performance	
				Analysts	Misc.
Number of Studies	4	7	5	5	4
Tests					
F.M.	0/4	0/7	0/5	2/5	0/4
R.V.	0/4	0/7	1/5	0/5	0/4
R.C.	0/4	0/7	-1/5	1/5	-1/4
CAT	0/4	1/7	1/5	0/5	1/4
AP	0/4	0/7	0/5	2/5	0/4
IDY	1/4	0/7	0/5	1/5	0/4
Con	0/4	-1/7	0/5	-1/5	0/4
N.O.	0/4	0/7	0/5	0/5	1/4
A.L.					
WA Training	0/4	0/7	0/5	0/5	0/4
WA Hazards	0/4	0/7	-1/5	-1/5	0/4
WA Analyze	0/4	0/7	0/5	0/5	0/4
WA Annoyances	0/4	0/7	0/5	0/5	0/4
WA Rewards	0/4	0/7	0/5	1/5	0/4
WA Soc. Resp.	0/4	0/7	0/5	0/5	0/4
WA Mech.	0/4	0/7	1/5	0/5	1/4
WA Supervisor	0/4	0/7	0/5	0/5	0/4
WA Phys. Dem.	0/4	0/7	0/5	-1/5	1/4,-1/4
WA Supervisee	0/4	0/7	0/5	0/5	0/4
WA Soc. Depr.	0/4	0/7	0/5	0/5	0/4
WA Undesirables	0/4	-1/7	1/5	0/5	1/5
WA Resourceful	1/4	0/7	0/5	-1/5	0/4
WA Tempo	0/4	0/7	1/5	0/5	1/4,-1/4
TTS Quick	0/4	0/7	0/5	1/5	0/4
TTS Physical	0/4	0/7	-1/5	1/5	1/4
TTS Outgoing	0/4	1/7	0/5	0/5	0/4
TTS Predominant	0/4	0/7	1/5	0/5	0/4
TTS Self-Conf.	1/4	0/7	0/5	0/5	0/4
TTS Solitary	0/4	0/7	0/5	0/5	0/4
TTS Question	0/4	0/7	0/5	0/5	0/4

The expectation has not been realized; the scores on the cognitive tests do not consistently correlate with these job performance criteria. In addition, the pattern of significant correlations makes neither logical nor psychological sense. For example, scores on the Reading Vocabulary test correlate positively with writing ability in one study and with oral communication in two of the studies, but scores on the Reading Comprehension test correlate negatively with writing ability in one study and with oral communication in one study. Since scores on the Reading Vocabulary test and the Reading Comprehension test usually have high positive correlations with each other, it is difficult to make sense out of the opposite signs for these two tests.

The next two criterion measures--independence/initiative and adaptability/ flexibility--would be expected to correlate most consistently with the scores on the work attitude and temperament scales. None of the scores on these scales was consistently and significantly related to these two criterion measures. The criterion measure of supervisory skill/managerial ability was used for 7 job groups and again no consistent pattern of correlations was found. Organizational ability was used as a criterion measure for 5 job groups and again the finding was the same--no consistent pattern of relationship.

The last 2 columns in Table 4 summarize the findings for overall job effectiveness for 5 categories of analyst jobs and for a miscellaneous group of 4 jobs--DCD IO's, [redacted]s, and Cable Analysts. Again there are no consistent relationships among the tests and ratings of overall job effectiveness.

STATSPEC

STATSPEC

In most of the 10 studies examined, the investigators have generated multiple regression equations to predict performance in the jobs studied. In one study [redacted] 1967), the investigator generated a multiple regression equation for predicting job performance of ORR MRA's even though he found no significant correlations between the single tests and scales on PATB and the criterion measures. This is not an acceptable procedure. [redacted] considered his 1967 study of MRA's a cross-validation of a 1958 study of ERA's done by [redacted] reported that the multiple regression equation for ERA's did not predict MRA job performance. Two of the 3 variables used to predict job performance of ERA's were weighted in the opposite direction for MRA's. He also reported that 10 interest scales from the Strong Vocational Interest Blank correlated significantly with job performance of MRA's but these were completely different from the 11 interest scales that correlated

STATINTL

STATINTL

STATINTL

significantly with job performance of ERA's. Despite the inconsistencies in the cross-validation data, [redacted] recommended that the separate regression equations be used to select applicants for ERA and MRA positions in the Agency. Such a recommendation is completely unacceptable. STATINTL

The only other studies that could be considered to be cross-validation studies were the two done using ODP computer programmers [redacted] 1978). Both of these studies used similar criterion measures but the multiple regression equations that were generated for predicting job performance were quite different in the two studies. Unfortunately [redacted] did not test out his multiple regression equations on the sample used by [redacted] and [redacted] STATINTL STATINTL STATINTL STATINTL

The correlational and multiple regression data presented in the 10 studies fail to meet the minimal standards of the APA 1/ and EEOC 2/. The standards that are most consistently and universally violated in the 10 studies are the following:

1. Criteria of job effectiveness are not based on systematic analyses of the jobs.

2. Samples used in the studies are inadequately described in relation to sex, race or ethnicity, age, educational levels and length of service.

1/ Op. Cit.

2/ Op. Cit. pp. 38304-7, Sec 15B.

3. Arithmetic means and standard deviations are not consistently reported.

4. Number of subjects reported in the study change over different variables with no explanation for the missing subjects.

5. Correlational data are incompletely reported. One study [redacted] 1958) used 14 criterion measures and one [redacted] 1973) used 8 criterion measures but reported correlations for only 5 criterion measures. Only 2 of the 10 studies reported complete correlational data.

STATINTL

STATINTL

6. Cross-validation has not been done. Cross-validation is particularly necessary when the number of predictors entering the study is greater than 4 or 5 and when the sample size is less than 200. None of the 10 studies have used samples of this size but all have used 30 or more predictors.

7. Negative scoring weights in regression equations should be used only if they have been verified by cross-validation in large samples. In all of the studies that have generated multiple regression equations, negative weights have been used and none have been cross-validated.

8. The samples used for validity study have not been representative of the applicant sample. The validity samples have been largely white males; females and minorities have been lacking or very much under-represented in the validity samples.

In view of the small number of subjects used in the validation studies, the lack of representation of females and minorities in the validation groups, the failure of the two attempts to cross-validate results, and the lack of cross-validation in general, none of the multiple regression equations that have been generated in the 10 studies should be used to predict job success or job placement. Responsible officials in the Agency should immediately institute whatever procedures are necessary to stop the operational use of the present multiple regression equations.

In the validity studies that have been done since 1970, a number of investigators have been using discriminant analyses in addition to or instead of multiple regression to determine whether scores on the tests and scales of PATB discriminate among employees who are rated high, average, or low in job performance. The use of this technique which requires that a sample of subjects that is too small to begin with be split into even smaller groups, some as small

as 5, is questionable at best. Elaborate equations have been generated on these inadequate samples and none has been cross-validated. The investigators who have used this technique have not reported their data adequately, particularly the full equations that have been generated. In one

STATINTL study [redacted] 1974) the investigator failed to predict job rating categories using the total group so he dropped the middle group and generated equations using only the top and bottom groups. This procedure is not acceptable and violates both APA and EEOC standards for validity studies.

In the few studies that have reported both correlational data and discriminant analysis data, the equations generated by discriminant analysis have used variables that show no significant correlations with the criteria of job performance. For example, in

STATINTL [redacted] study of ODP computer programmers (1978), the equations generated by discriminant analysis included many work attitude and temperament variables but only two of these correlated significantly with job performance ratings. Another troublesome aspect of the discriminant analysis is that the numbers in the different categories of rated job performance tend to be very unequal, therefore the reported accuracy of the discriminant analysis in assigning individuals to different categories is extremely misleading

since the base rates vary enormously.

The inconsistency between the correlational and the discriminant analysis data, the extremely small size of the subgroups used in the discriminant analyses, the differences in base rates for the different categories of rated job performance, and the lack of cross-validation of the equations generated from discriminant analysis indicate that these equations should not be used to predict job performance or to recommend placement in particular jobs. Responsible officials in the Agency should immediately stop the use of these equations for these purposes.

In reviewing the validity studies and the memoranda prepared by psychologists in PSS, we have been extremely troubled by the unrestrained enthusiasm with which the psychological staff has promoted the operational use of PATB test scores for selection and placement of personnel in the Agency. The enthusiastic promotion of these uses bears no relationship to the adequacy of the data. Although the majority of the investigators explicitly stated in their studies that the sample used was small, that the number of significant correlations did not exceed what one would expect by chance, and that cross-validation was needed, they then promote the use of the data in a manner that implies a degree of accuracy that is not supported by the evidence.

There are three parts of PATB for which little or no validity data are available. These are the Biographical Information Blank, the writing sample and the Strong-Campbell Interest Inventory. The Biographical Information Blank was apparently used in a number of the studies but only one study [redacted] 1974) reported data on it. Like all the other tests in the battery, there were no consistent correlations between these scores and ratings of job performance.

STATINTL

The writing sample has not been validated at all, probably because it is scored impressionistically rather than quantitatively. The lack of attention given to establishing the validity of the writing sample is inexplicable given the importance of writing ability for many of the jobs in the Agency.

No validity studies have been done on the Strong-Campbell Interest Inventory. It appears as though the psychologists have assumed that the validity data accumulated for the old Strong-Vocational Interest Blank, which is, by the way, completely inadequate and unconvincing, can be applied to the Strong-Campbell Interest Inventory (SCII).

It cannot be; the SCII is a new instrument with quite different characteristics from the old Strong Vocational Interest Blank. We would recommend that no reports of scores on the SCII be made to units until the validity of these scores has been established.

In closing this section on criterion-related validity, we would like to make a few additional comments. First, our review of the criterion-related validity of PATB has been greatly hindered by the inadequacies and general poor quality of the studies made available to us. We have mentioned some of the inadequacies previously--the lack of complete data particularly standard deviations and complete correlational matrices, the failure to describe the characteristics of the validation sample, and in 6 studies the complete omission of all statistical data. The Chief of PSS told us that the reports were written to be used by people in the units who were naive in testing and statistics. Evidently no separate technical report that would provide complete data was kept.

Second, we are seriously troubled by the fact that the samples used in the validation studies, as far as we can determine from the inadequate reporting, are primarily or

solely composed of white males. To the extent that this is true, then the use of multiple regression equations or discriminant analysis equations based predominantly or solely on white males are potentially unfair to women and minorities. We recognize that, with the small number of persons in each job category, it is impossible to do separate validity studies for white males, females and minorities. However, we think that the lack of representation of women and minorities in the groups studied and the general overall weakness of the validity data indicate that it would be wise to stop using the equations for selection and placement.

Third, there does not seem to have been in the past or now any systematic plan for validating PATB. The tests have been used in the Agency since the late 1950's, a period of 20 years at least. Despite this length of time, we could find only 10 studies in which at least minimum requirements for data reporting were met. Perhaps the lack of attention given to establishing the validity of PATB has been due, in part at least, to the lack of cooperation from the units in the Agency. It is our impression that psychologists in PSS have done studies on the validity of PATB for a particular

job when a unit has requested it, not on the basis of any overall plan for validating the test battery. In order to have good procedures for selecting personnel, the authority of responsible people in the Agency must back up and support the efforts of the psychological staff to obtain the required cooperation. We could find no evidence of such support in the Agency. In part the lack of any systematic plan for establishing the validity appears to be due to the fact that no one appears to have been assigned the primary responsibility for doing it and no one has been given the resources to do it.

Fourth, we have been disturbed by the fact that the Agency has been looking at its personnel selection procedures on a piece-meal basis. We have been assigned the task of examining PATB and its role in personnel selection and placement. However, PATB is only one element of the job selection procedures. It is a factor in selection and placement for only about two-thirds of the applicants for professional positions and only after one of the units indicates an interest in the applicant. We have not been able to determine to our satisfaction how important a role that performance on PATB plays in making employment decisions.

We have expressed serious doubts about the adequacy of the validity evidence on PATB and would express even more serious doubts about the other procedures for which no data on validity exist. The Agency has expressed concern about bias or unfairness in PATB but it should be even more concerned about bias or unfairness in other selection procedures since the initial opening of the gate for employment is based on information other than PATB.

Fifth, the validation studies of PATB appear to be mechanistic, atheoretical, not based on an adequate analysis of jobs and completely divorced from other procedures or information used to select people for jobs. In all of the studies that we reviewed the procedure used was to throw all the variables in, then examine what came out. The question as to whether the findings made any psychological sense in relation to the job being studied was not asked. Although there were numerous opportunities to study the utility of PATB; i.e. the improvement in accuracy of selecting satisfactory employees when PATB was used, no such study was made. Whether a test should be used for personnel selection when other information is readily available depends not on the validity of the test but on its incremental validity; that is, what it adds to the soundness of the judgments that would otherwise be made.

Sixth and last, there appears to be no systematic plan for reviewing proposed validity studies critically and for supervising the quality of the studies or the adequacy of the written report of the study. In view of the poor quality of the majority of the studies that we examined, some form of quality control is needed.

Validity of PATB for Foreign Language Training

We reviewed four studies that reported data on the relationship between scores on PATB and success in foreign language training. [redacted] STATINTL

STATINTL [redacted] January 1977 and December 1977). The only two tests in PATB that correlated with any consistency with success in foreign language training were Reading Vocabulary and Reading Comprehension but these consistently related to foreign language training only in French and Spanish. One would have to conclude that the scores on PATB provide very little help in predicting success in all? foreign language training.

Reliability Studies of PATB

Little attention has been given to determining the reliabilities of the scores from PATB. The Test Data Book No. 15 dated 1 July 1958 reports internal consistency

reliabilities for some of the cognitive tests included in the battery at that time. This source gives no reliability data for the Considerations test or the Numerical Operations test. The reliabilities for males range from .80 for the Interpretation of Data test to .91 for the Reading Comprehension test. For females, the reported reliabilities range from .75 for the Interpretation of Data test to .87 for the Reading Comprehension test. No reliability data for females are reported for the work attitude scales. For 75 males, test-retest reliabilities range from .49 for WA Supervisee to .81 for WA Mechanical with a median reliability of .71. The time interval between administration of the two tests was not reported. The reliabilities of the scores on the temperament scales for males computed by Kuder-Richardson #20 ranged from .57 for TTS Quick to .86 for TTS Predominant with a median reliability of .68. For females, the reliabilities ranged from .46 for TTS Quick to .82 for TTS Predominant with a median reliability of .70.

Only one study of the reliability of the scores on PATB has been done since 1958. [redacted] (1975) studied the test-retest reliabilities of the scores on the eight cognitive tests of PATB. The number of subjects used to compute

STATINTL

reliabilities varied across tests. For males, the numbers ranged from 276 on the Contemporary Affairs Test to 337 for the Figure Matrices, Reading Vocabulary and Reading Comprehension. For females, the number ranged from 50 on the Contemporary Affairs Test to 80 for Reading Vocabulary and Reading Comprehension. He reports that the interval between tests ranged from several months to more than 10 years, but he does not report how many subjects fell into the different time periods. He gives no other data on the subjects, but the Chief of PSS told us that these subjects were drawn from a pool of people who were employed by the Agency and who had come or been sent to PSS because they were either performing poorly on their jobs or were having personal problems. Because of the nature of the sample and the inadequacies in reporting, we have considerable difficulty in making value judgments about the reliabilities. For men, the test-retest correlations ranged from .56 for the Considerations test to .86 for Reading Comprehension with a median reliability of .76. For females, the reliabilities ranged from .46 for the Contemporary Affairs test to .85 for the Reading Vocabulary test with a median reliability of .75. How much of the instability in scores on the tests between

the two testing dates is due to the tests themselves and how much is due to the non-representative sample used to compute them is impossible to determine.

STATINTL [redacted] also reported split-half reliabilities for six of the same cognitive tests using a sample of 195 males randomly selected from the 1973-1974 male applicants. For this sample, the split-half reliabilities corrected for full length by the Spearman-Brown Prophecy Formula ranged from .77 for the Interpretation of Data test to .88 for the Reading Vocabulary test with a median reliability of .86. The reliability of .88 for scores on the Reading Vocabulary test is inflated because it is a speeded test. The reliability of .78 for the Reading Comprehension test and the reliability of .77 for the Interpretation of Data test are also inflated because the items on these tests are not completely independent.

No standard errors of measurement of the tests and scales of PATB have been computed for an applicant sample. This is a serious oversight. We suspect that the standard errors of measurement are likely to be relatively large in light of the limited reliabilities of the tests and particularly so for the work attitude scales and temperament scales whose reliabilities are much lower than those for the cognitive tests.

Overall, the reliabilities of the test and scales of PATB for males are not impressive. The reliabilities for the work attitude scales appear to be much too low to justify the emphasis given them in the equations that are used to predict job success. We are not sure that the temperament scales for which we have reliability data are the same ones that are currently being used. If they are not, then there are no reliability data for the current scales. If they are, the reliabilities of 4 of the 6 scales are much too low to use the scores for selection decisions.

Reliabilities of the cognitive tests for females are inadequate. Three of the tests, Figure Matrices, Contemporary Affairs Test and Considerations, have reliabilities below .60 which are completely inadequate for making decisions about individuals. Only the Reading Vocabulary, Arithmetic Problems and Numerical Operations tests have reliabilities in the .80's which would be considered to be adequate for making decisions about individuals. The other cognitive tests--Reading Comprehension and Identification of Data--have reliabilities in the .70's which would be considered marginal. No reliability data of any kind is available for females on the work attitude scales. In view

of the importance given to these scales by the psychological staff in making recommendations for employment, the lack of any reliability data is a serious matter. The use of these scores for women should be discontinued until their reliabilities are determined. Reliability data for women on the temperament scales are available for only 5 of the 6 scales and they tend to be too low on 4 out of the 5 for individual decisions.

No reliability data are available for minorities. Again, this is a serious oversight and needs to be corrected.

deflammatory and incorrect

Summary and Conclusions

We have reviewed 23 studies that purported to present evidence on the validity and reliability of PATB. Six studies had to be discarded because no data were reported. Two other studies had to be discarded because they reported no data on PATB. One study of black and white employees had to be discarded because of deficiencies in statistical analyses of the data. After discarding these studies, we were left with 14 studies, 4 of which were related to foreign language training and 10 to job performance in various units of the Agency. All of the evidence for validity of PATB is found in these 14 studies and, on the whole, it is weak and unconvincing.

?

Rubbish. There are data reported in each of the studies they choose to disregard

Our major findings in relation to the validity of PATB are as follows:

1. No consistent pattern of correlations for similar jobs in the Agency or for similar criteria of job performance has been found. The criterion-related validity of PATB still needs to be determined.

2. Equations generated through multiple regression analyses and discriminant analyses have been based on extremely small samples and have not been cross-validated. They should not be used to select or place personnel in Agency jobs.

3. There is no evidence that the initial construction of PATB was based on a systematic job analysis; therefore the content and construct validity of PATB has not been demonstrated.

4. The samples used to study the criterion-related validity of PATB have been composed solely or primarily of white males. There are no validity data for females or minorities.

5. No validity data are available for the Biographical Information Inventory, the writing sample, and the Strong-Campbell Interest Inventory.

ILLEGIB

[redacted]

6. None of the tests or scales in PATB has shown any consistent significant correlations in predicting foreign language training. The battery appears to be useless for this purpose.

7. The evidence on validity presented in the 14 studies does not meet the minimal standards for validity set by APA or EEOC.

Evidence on the reliability of the separate tests and scales of PATB is scanty. Evidently little attention has been paid to this extremely important aspect of the tests. On the basis of the limited amount of information available to us concerning the reliabilities of the separate tests, our findings are as follows:

1. For white males, only the tests of Reading Comprehension, Arithmetic Problems, and Numerical Operations have high enough reliabilities to be used to make decisions about individuals. The reliabilities of the Reading Vocabulary and Identification of Data tests are marginal. The reliabilities of the Figure Matrices, Contemporary Affairs, and Considerations tests are unacceptable; they are all below .70.

2. The reliabilities of the work attitude scales for white males are generally unacceptable. Only two of the scales have reliabilities of .80 or higher and 5 have reliabilities below .70.

3. The reliabilities of the temperament scales for white males are unacceptable. Only one has a reliability of .80 or higher, and three have reliabilities below .70.

4. For white females, only three of the cognitive tests have acceptable reliabilities--Reading Vocabulary, Arithmetic Problems, and Numerical Operations. Two tests--Reading Comprehension and Identification of Data--have marginal reliabilities. The reliabilities of Figure Matrices, Contemporary Affairs, and Considerations are unacceptable.

5. There are no reliability data for white females on the work attitude scales.

6. The reliabilities of the temperament scales for women, in general, are unsatisfactory. No reliability data are reported for 1 scale. Two scales have reliabilities below .70 and only 1 has a reliability higher than .80.

7. No reliability data are available for the writing sample.

8. No reliability data are available for minorities.

In general we found the quality of reporting in the studies poor. Samples were inadequately described in all of the studies. Data were incompletely reported. Conclusions drawn in the majority of the studies were not supported by the data presented in the study. Most distressing to us was the enthusiasm shown by the investigators to encourage uncritical use of the results of the study when the results of the study truly did not support such use.

References

Equal Employment Opportunity Commission (EEOC) Uniform Guidelines for employee selection procedures. Federal Register, August 25, 1978, 43 (166), 38290-38315.

Standards for Education And Psychological Tests. Washington, D.C.: American Psychological Association, 1974.

ADMINISTRATIVE-INTERNAL USE ONLY

Approved For Release 2002/01/25 : CIA-RDP00-01458R000100130010-7

Next 2 Page(s) In Document Exempt

Approved For Release 2002/01/25 : CIA-RDP00-01458R000100130010-7